PROJECTE DE SÈRIES TEMPORALS

ANÀLISI DE DADES

Grau en ciència i enginyeria de dades.

Facultat d'informàtica de Barcelona (FIB)

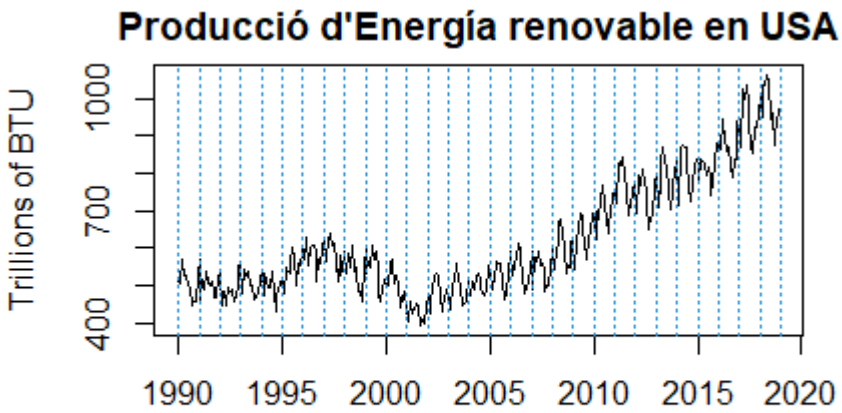
Universitat Politècnica de Catalunya (UPC)

Santi Sayol Pruna

Roger Bel Clapés

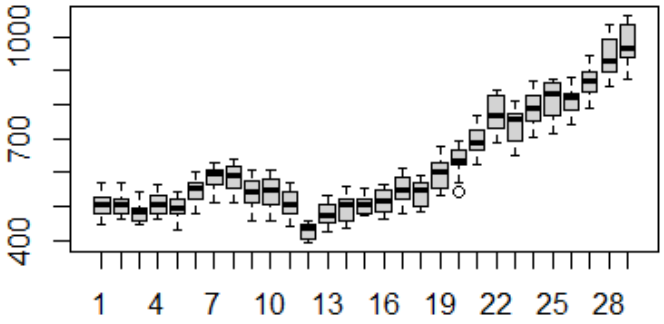
INDENTIFICACIÓ:

Per començar carreguem el fitxer de dades amb el que treballarem i el definim com una "time series", en el nostre cas hem escollit el fitxer "RenewUSA.dat" que recull informació sobre la producció total d'energia de tipus renovable als Estats Units entre el 1990 i el 2019 en trilions de BTU (British Thermal Units). Si representem aquestes dades obtenim la gràfica següent:

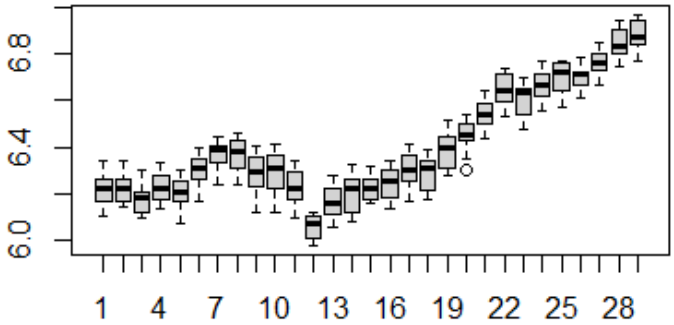


Representació de la sèrie original.

Visualitzant el "boxplot" de les dades observem que la seva variància no és constant i que no presenta volatilitat, així que apliquem la transformació logarítmica a la sèrie original per aconseguir que la variància si que sigui constant.

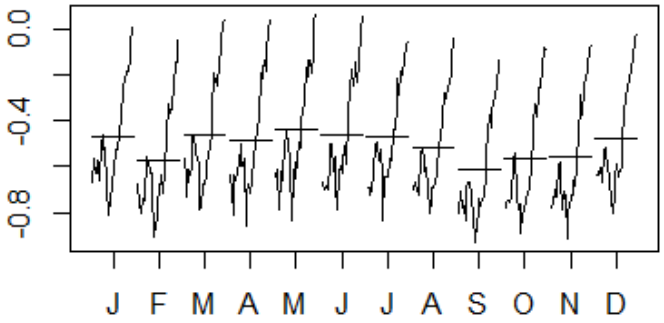


"Boxplot" de la sèrie original.

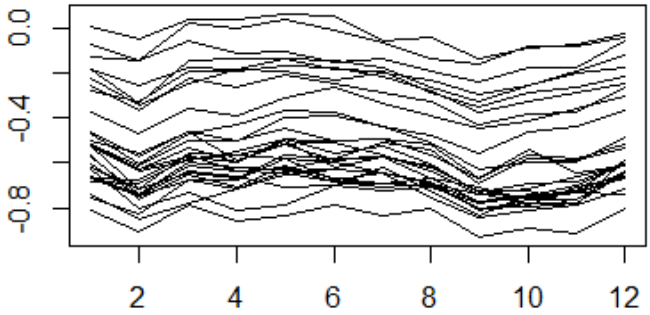


"Boxplot" de la sèrie després d'aplicar el logaritme.

Continuem representant el "monthplot" i les dades de cadascun dels anys de la sèrie amb la transformació logarítmica per separat en un mateix gràfic utilitzant el "ts.plot()" per a poder detectar si la sèrie presenta estacionalitat.

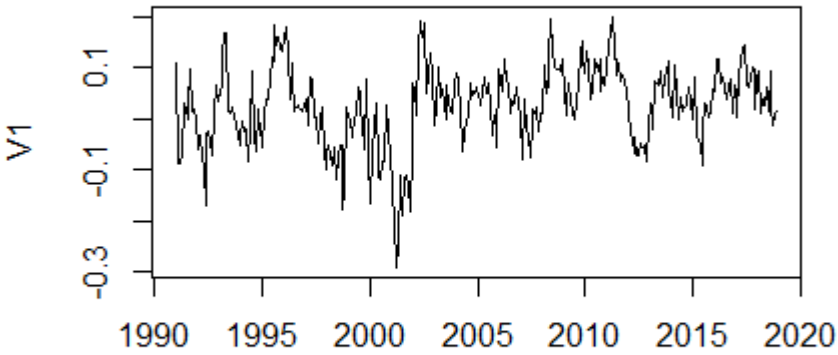


"Monthplot"



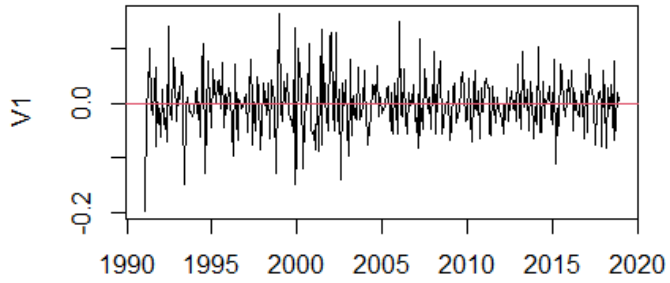
"ts.plot()"

S'observa que la sèrie presenta estacionalitat, per tant apliquem diferenciació estacional per a eliminar-la.

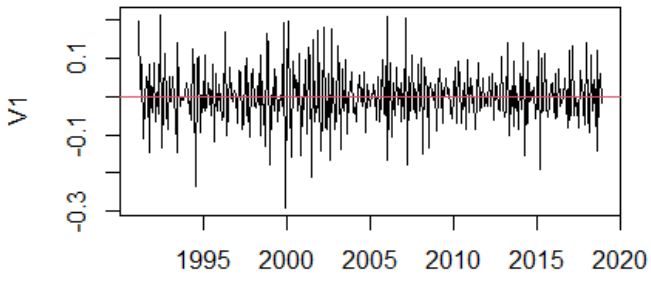


Representació de la sèrie després d'aplicar el logarimet i una transformació estacinal a la sèrie original.

En aquest punt la nostra sèrie ja presenta variància constant i no té estacionalitat, però en el gràfic anterior es pot apreciar que la mitjana no és constant, així que apliquem diferenciació regular dues vegades i posteriorment descartarem una d'aquestes diferenciacions en cas que sigui innecessària basant-nos en l'anàlisi del valor de la variància de la sèrie després de cada transformació que li hem aplicat.



Sèrie amb una diferenciació regular



Sèrie amb dos diferenciacions regulars

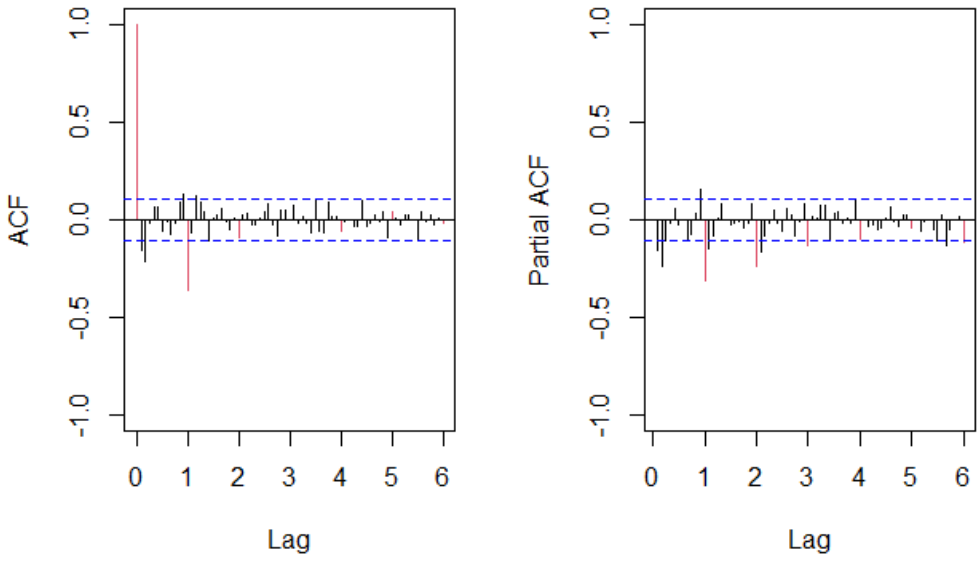
Ara clarament la mitjana de la sèrie és constant, però com ja hem comentat hem de comprovar si realment era necessari aplicar dues diferenciacions regulars o només una era suficient. Analitzem el valor de la sèrie després de cada una de les transformacions que hi hem aplicat:

```
             v1
v1 24119.97
             v1
v1 0.05493051
             v1
v1 0.006061437
             v1
v1 0.002723263
             v1
v1 0.006207768
```

Variància de la sèrie després de cada transformació que hem aplicar.

S'observa que la variància disminueix després de cada transformació, fet que s'espera al intentar transformar una sèrie en estacionària, però que després de la segona diferenciació regular aquesta augmenta notablement, fins i tot supera el valor de la variància abans d'aplicar-hi la primera transformació regular. Basant-nos en aquest fet i que en les gràfiques on es representa la sèrie després de cada diferenciació regular es pot identificar que la mitjana ja sembla constant en la primera de les dues imatges finalment descartem la segona diferenciació regular.

D'aquesta manera ja hem aconseguit transformar la nostra sèrie en estacionària, després d'aplicar una transformació logarítmica, una diferenciació estacional i una diferenciació regular. Ara desem aquesta sèrie ja estacionaria amb el nom de "w_serie" i representem el seu ACF i PACF:



ACF i PACF de la sèrie estacionaria

En la part estacional podem observar que en el PACF pràcticament tots els retards són significatius, mentre que en el ACF només el primer retard és significatiu, tot i que el segon dista molt poc de les bandes de confiança. Per aquest motiu, per simplificar els models, plantejarem dos models amb part estacional ARMA(0,2)$_{12}$ i en cas que el segon coeficient de la part estacional realment no sigui significatiu en el model l'eliminarem.

Per altra banda, en la part regular en el ACF l'últim retard que es considera estadísticament no nul és el segon i en el PACF és el tercer. D'aquesta manera els dos models que ajustarem per a la nostra sèrie seran un ARIMA(0,1,2)(0,1,2)$_{12}$ i un ARIMA(3,1,0)(0,1,2)$_{12}$.

ESTIMACIÓ:

-Model ARIMA(0,1,2)(0,1,2)$_{12}$:

Comencem estimant el nostre model sobre la sèrie ja estacionaria "w_serie":

```
Call:
arima(x = w_serie, order = c(0, 0, 2), seasonal = list(order = c(0, 0, 2), period = 12))

Coefficients:
          ma1      ma2     sma1     sma2  intercept
      -0.2442  -0.2300  -0.7281  -0.1604      2e-04
s.e.   0.0545   0.0565   0.0590   0.0616      2e-04

sigma^2 estimated as 0.001573:  log likelihood = 597.04,  aic = -1182.08
```

```
T-ratios: -4.48 -4.07 -12.33 -2.6 0.74
```

Veiem que la mitjana no és significativa, així que tornem a estimar el model sobre la sèrie amb només la transformació logarítmica:

```
Call:
arima(x = logserie, order = c(0, 1, 2), seasonal = list(order = c(0, 1, 2),
    period = 12))

Coefficients:
          ma1      ma2     sma1     sma2
      -0.2423  -0.2262  -0.7235  -0.1579
s.e.   0.0547   0.0560   0.0586   0.0615

sigma^2 estimated as 0.001579:  log likelihood = 596.76,  aic = -1183.51

T-ratios: -4.43 -4.04 -12.34 -2.57
```

Ara tots els coeficients són significatius, així que ens quedem amb el model ARIMA(0,1,2)(0,1,2)$_{12}$.

-Model ARIMA(3,1,0)(0,1,2)$_{12}$.

De nou comencem estimant el nostre model sobre la sèrie ja estacionaria “w_serie”:

```
Call:
arima(x = w_serie, order = c(3, 0, 0), seasonal = list(order = c(0, 0, 2), period = 12))

Coefficients:
          ar1      ar2      ar3     sma1     sma2  intercept
      -0.2432  -0.2763  -0.1111  -0.7160  -0.1583      1e-04
s.e.   0.0563   0.0547   0.0552   0.0587   0.0628      3e-04

sigma^2 estimated as 0.00158:  log likelihood = 596.97,  aic = -1179.93


T-ratios: -4.32 -5.05 -2.01 -12.19 -2.52 0.58
```

De nou la mitjana no és significativa i hem de tornar a estimar el model sobre la sèrie amb només la transformació logarítmica:

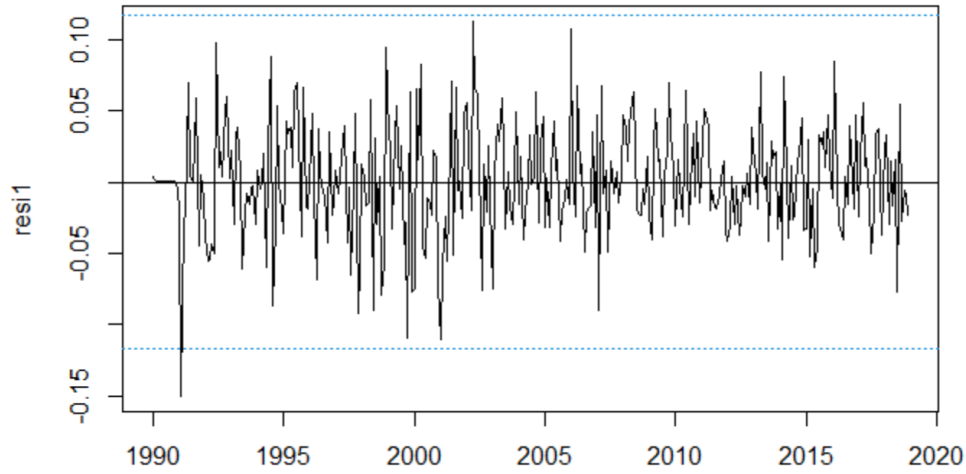
```
Call:
arima(x = logserie, order = c(3, 1, 0), seasonal = list(order = c(0, 1, 2),
    period = 12))

Coefficients:
          ar1      ar2      ar3     sma1     sma2
      -0.2428  -0.2754  -0.1094  -0.7133  -0.1563
s.e.   0.0564   0.0548   0.0551   0.0586   0.0627

sigma^2 estimated as 0.001584:  log likelihood = 596.79,  aic = -1181.58


T-ratios: -4.31 -5.03 -1.98 -12.18 -2.49
```

Si ens centrem de manera estricta al criteri de només considerar un coeficient com a significatiu si el seu “T-ratio” és superior a dos en aquest cas el tercer coeficient de la part regular no ho seria, però el seu valor és molt proper a dos i en cas de que no el considerem el model disminueix el seu aic, per aquest motiu finalment ens quedem amb el model ARIMA(3,1,0)(0,1,2)$_{12}$.

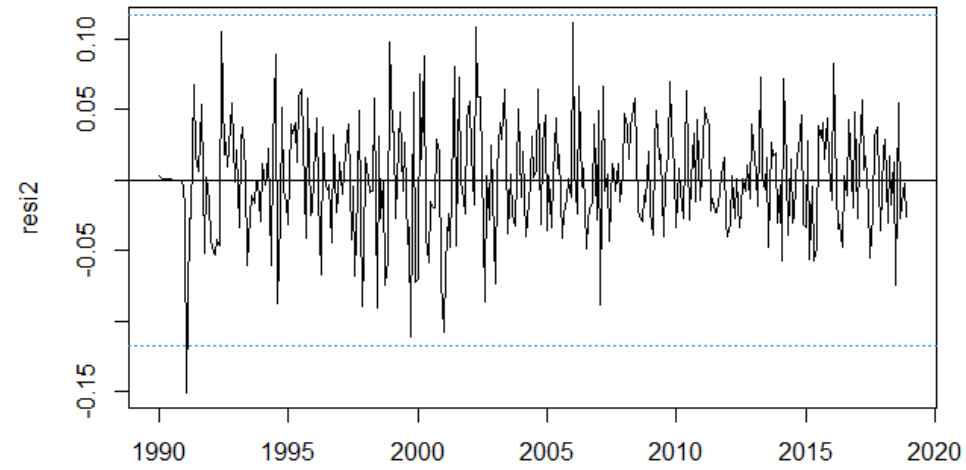
VALIDACIÓ:

Anàlisi complet dels residus:

Si representem gràficament els residus es pot observar que per els dos models pràcticament tots els punts cauen dintre de les bandes de confiança, tot i que hi ha algunes excepcions puntuals, probablement això és degut a la presència de "outliers" entre les dades.



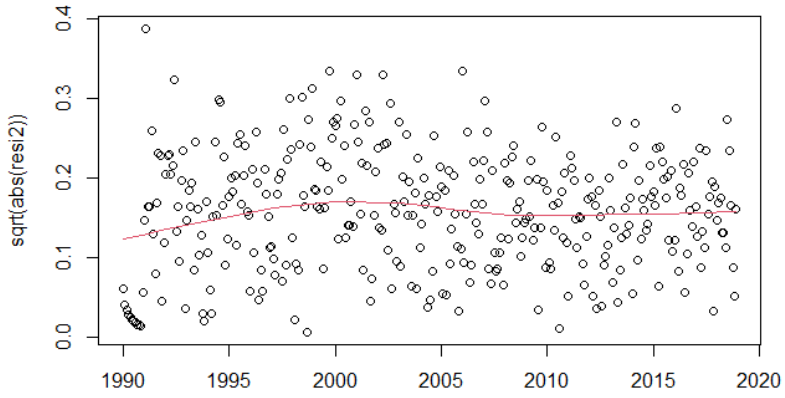
Representació dels residus del model ARIMA(0,1,2)(0,1,2)$_{12}$



Representació dels residus del model ARIMA(3,1,0)(0,1,2)$_{12}$

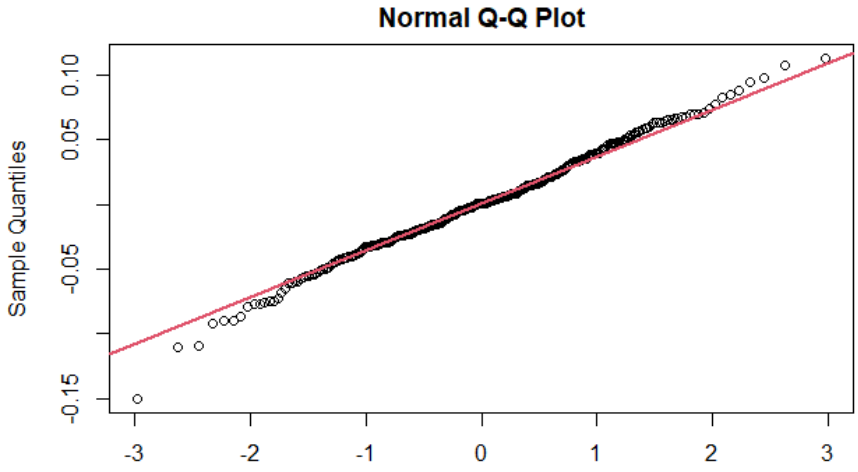
Analitzant l'estimació de la variància residual al llarg del temps s'identifica en ambdós casos una tendència creixent al principi. Així que no es pot considerar que la variància sigui del tot constant.

Arrel quadrada del valor absolut dels residus amb ajust suau del model *ARIMA(0,1,2)(0,1,2)$_{12}$*

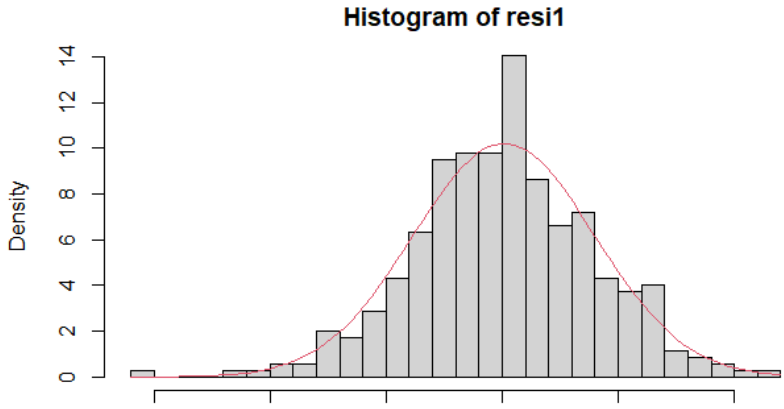


Arrel quadrada del valor absolut dels residus amb ajust suau del model *ARIMA(3,1,0)(0,1,2)*$_{12}$

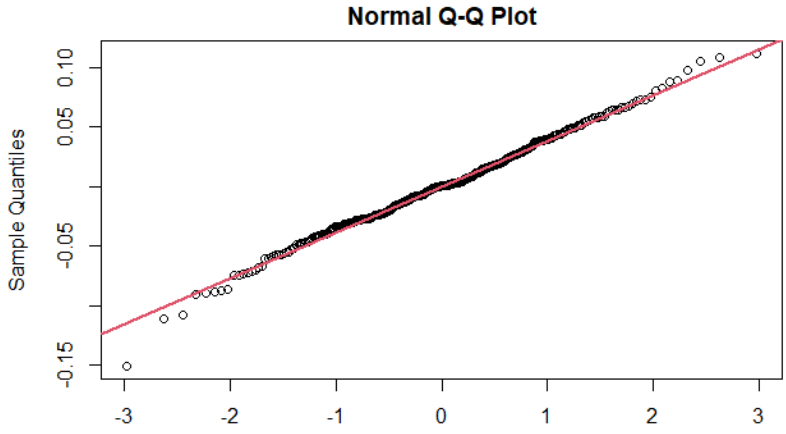
Observant el "plot de normalitat" i l'histograma amb la corba normal superposada podem veure que en general els residus dels dos models s'ajusten prou bé a una distribució normal, tot i que a les cues hi ha alguns valors que disten lleugerament del seu valor esperat.



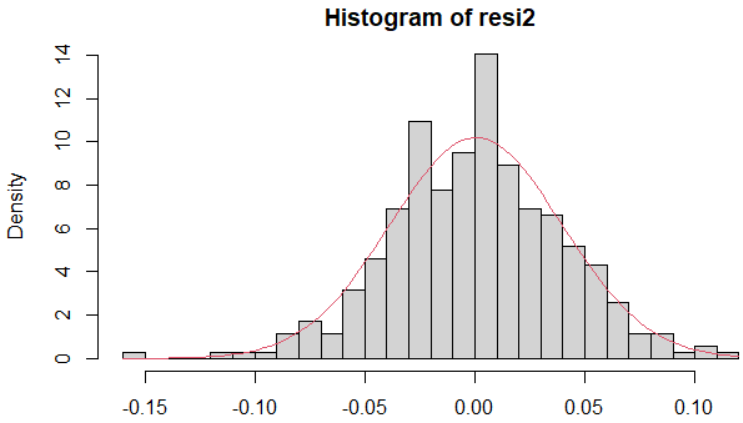
Plot de normalitat dels residus model *ARIMA(0,1,2)(0,1,2)*$_{12}$



Histograma amb la corba normal superposada dels residus model *ARIMA(0,1,2)(0,1,2)*$_{12}$



Plot de normalitat dels residus model *ARIMA(3,1,0)(0,1,2)*$_{12}$



Histograma amb la corba normal superposada dels residus model *ARIMA(3,1,0)(0,1,2)*$_{12}$

Si realitzem el test de Shapiro-Wilks als residus els dos "p-values" són superiors a 0.05 i per tant podem confirmar la hipòtesi de normalitat per els residus dels dos models.

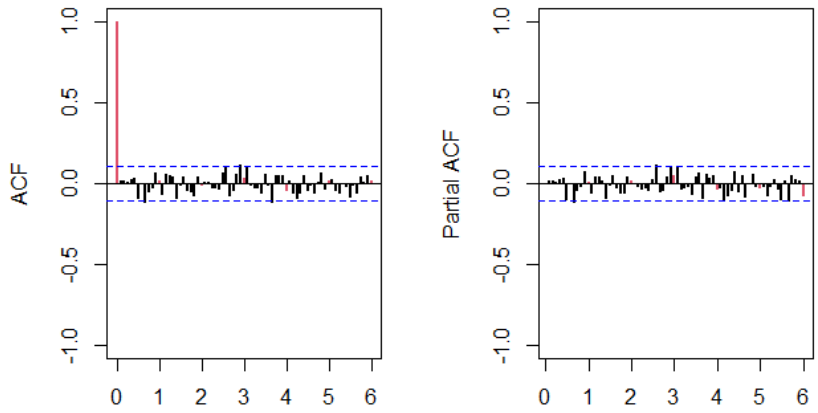
```
        Shapiro-Wilk normality test

data:  resi1
W = 0.9941, p-value = 0.1973
```

Test de Shapiro-Wilks del model ARIMA(0,1,2)(0,1,2)$_{12}$

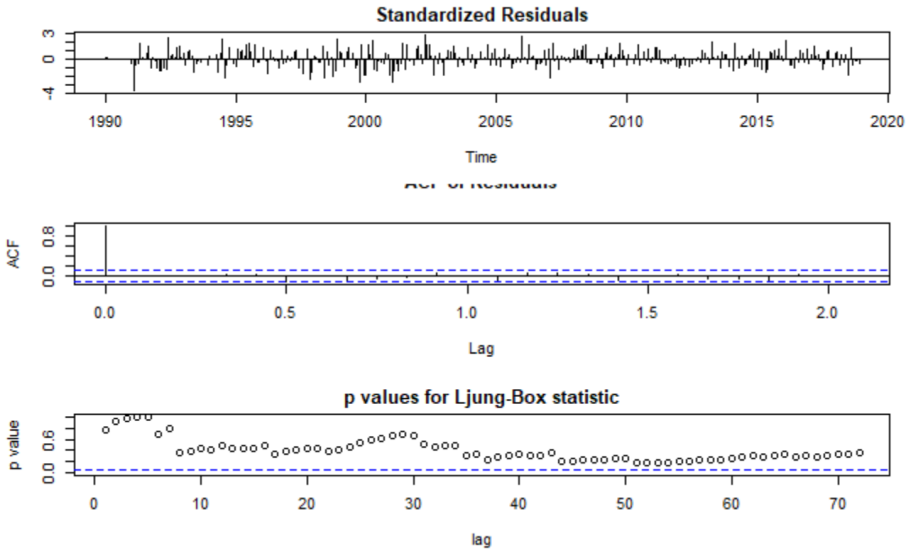
```
        Shapiro-Wilk normality test

data:  resi2
W = 0.99349, p-value = 0.1387
```

Test de Shapiro-Wilks del model ARIMA(3,1,0)(0,1,2)$_{12}$

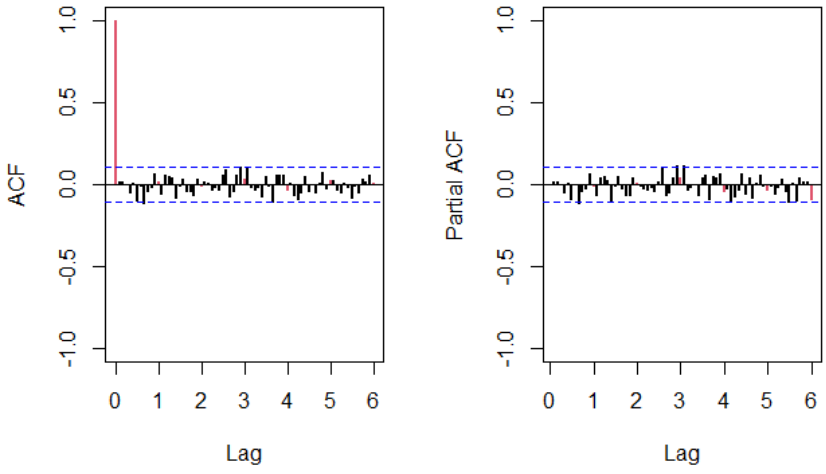
Per últim, representem el ACF i el PACF dels residus del dos models i apliquem també el test de Ljung-Box i analitzant els resultats obtinguts podem concloure que els residus són independents.



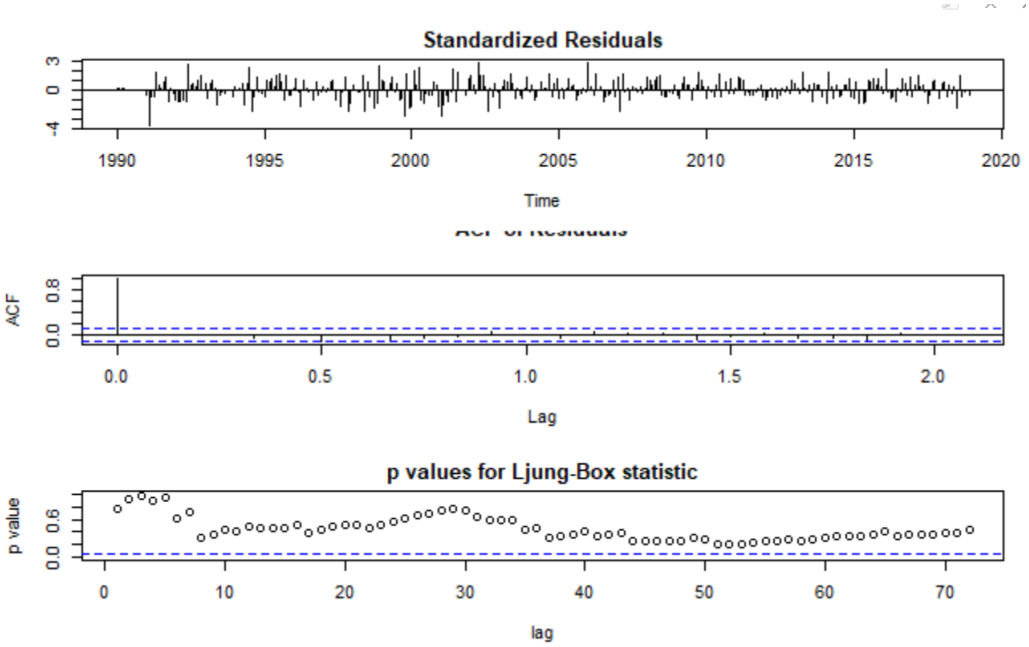
ACF i PACF dels residus del model ARIMA(0,1,2)(0,1,2)$_{12}$



Resultats del test de Ljung-Box dels residus del model *ARIMA(0,1,2)(0,1,2)$_{12}$*



ACF i PACF dels residus del model ARIMA(3,1,0)(0,1,2)$_{12}$



Resultats del test de Ljung-Box dels residus del model *ARIMA(3,1,0)(0,1,2)*$_{12}$

Un cop acabat el anàlisi dels residus, calculem el mòdul de les arrels dels polinomis característics de la part AR i MA dels dos models i observem que totes són superiors a 1, per tant, podem concloure que en ambdós casos es tracta de models causals i invertibles.

Calculem les mesures d´adequació a les dades AIC i BIC i obtenim els següents resultats:

-AIC model ARIMA(0,1,2)(0,1,2)$_{12}$ = -1183.513

-BIC model ARIMA(0,1,2)(0,1,2)$_{12}$ = -1161.442

-AIC model ARIMA(3,1,0)(0,1,2)$_{12}$ = -1181.577

-BIC model ARIMA(3,1,0)(0,1,2)$_{12}$ = -1158.692

En els dos casos els AIC són molt similars, el que indica que els dos models ajusten les dades de forma correcta. Per altra banda, el BIC és lleugerament superior pel primer model, això es deu al fet que aquest estimador penalitza de forma més estricta els models amb complexitat elevada.

Ara per verificar l'estabilitat dels models i avaluar la seva capacitat de predicció reservem les últimes 12 observacions de la nostre sèrie i tornem a ajustar els dos models:

```
Call:
arima(x = lnserie2, order = c(0, 1, 2), seasonal = list(order = c(0, 1, 2),
    period = 12))

Coefficients:
          ma1      ma2     sma1     sma2
      -0.2255  -0.2389  -0.7235  -0.1557
s.e.   0.0558   0.0574   0.0600   0.0620

sigma^2 estimated as 0.001596:  log likelihood = 573.47,  aic = -1136.94
```

Estimació del model ARIMA(0,1,2)(0,1,2)$_{12}$ reservant les últimes 12 observacions.

```
Call:
arima(x = lnserie2, order = c(3, 1, 0), seasonal = list(order = c(0, 1, 2),
    period = 12))

Coefficients:
          ar1      ar2      ar3     sma1     sma2
      -0.2259  -0.2766  -0.1051  -0.7130  -0.1541
s.e.   0.0576   0.0557   0.0563   0.0599   0.0634

sigma^2 estimated as 0.001602:  log likelihood = 573.32,  aic = -1134.63
```

Estimació del model ARIMA(3,1,0)(0,1,2)$_{12}$ reservant les últimes 12 observacions.

Com que si comparem els dos models ajustats reservant les últimes 12 observacions amb els models originals tots els coeficients tenen el mateix signe, la mateixa magnitud, segueixen sent significatius i tenen un AIC semblant el model es manté estable.

Per últim, avaluem la capacitat de previsió dels models calculant el seu RMSE, MAE, RMSPE, MAPE:

-RMSE model ARIMA(0,1,2)(0,1,2)$_{12}$ = 27.315

-MAE model ARIMA(0,1,2)(0,1,2)$_{12}$ = 23.40622

-RMSPE model ARIMA(0,1,2)(0,1,2)$_{12}$ = 0.02819733

-MAPE model ARIMA(0,1,2)(0,1,2)$_{12}$ = 0.02406141

-RMSE model ARIMA(3,1,0)(0,1,2)$_{12}$ = 26.82785

-MAE model ARIMA(3,1,0)(0,1,2)$_{12}$ = 22.83839

-RMSPE model ARIMA(3,1,0)(0,1,2)$_{12}$ = 0.02754265

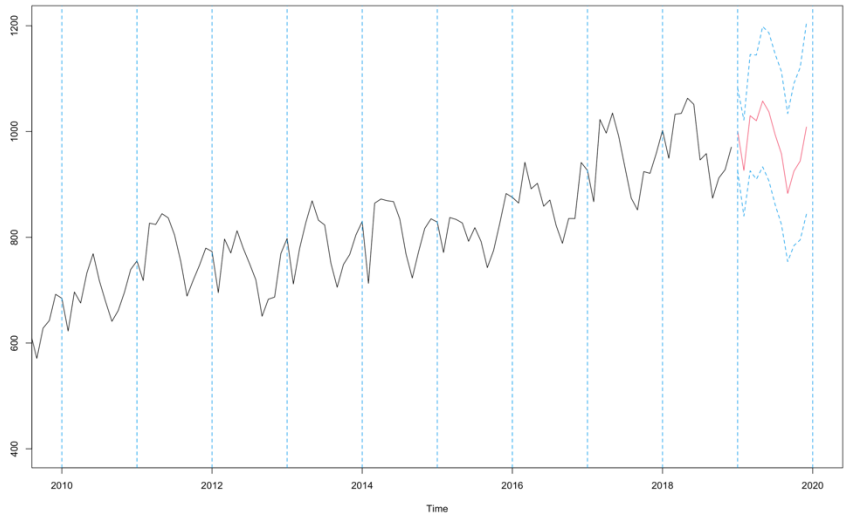
-MAPE model ARIMA(3,1,0)(0,1,2)$_{12}$ = 0.02344904

Com a conclusió, els dos models ajustats tenen característiques molt similars com s'ha pogut observar en l'anàlisi dels residus i en general les mesures de la seva capacitat de previsió també són molt semblants. Tot i així, ja que hem d'escollir un dels dos per a realitzar els últims dos apartats d'aquesta pràctica, seleccionarem el model ARIMA(3,1,0)(0,1,2)$_{12}$ per a dur a terme les prediccions basant-nos en que el seu RMSE,

MAE, RMSPE i MAPE són lleugerament inferiors, tot i que en aquest cas sembla que seria irrellevant quin dels dos models triem ja que tenen característiques tan similars que tot sembla indicar que es comportaran de forma molt semblant a l'hora de realitzar prediccions.

PREVISIONS:

Amb el model ARIMA(3,1,0)$(0,1,2)_{12}$ hem obtingut les següents previsions a llarg termini per als dotze mesos posteriors a la última observació, és a dir, per a tot l'any 2019 fins a 2020. De color blau s'observen els intervals de confiança i en vermell la predicció:



Per a aquest model (sense tractament d'atípics) l'amplada mitjana dels intervals de confiança és de 265,5365. Aquesta dada ens servirà posteriorment per a veure si després del tractament d'atípics es veu reduïda i per tant millora la predicció.

TRACTAMENT D'ATÍPICS:

Detecció automàtica:

Quan executem la detecció automàtica d'atípics obtenim la següent taula, on s'hi pot observar la data de l'atípic detectat, per a poder ubicar-lo entre tota la sèrie temporal, l'efecte que aquest ha tingut en la sèrie i el tipus (TC, LS o AO):

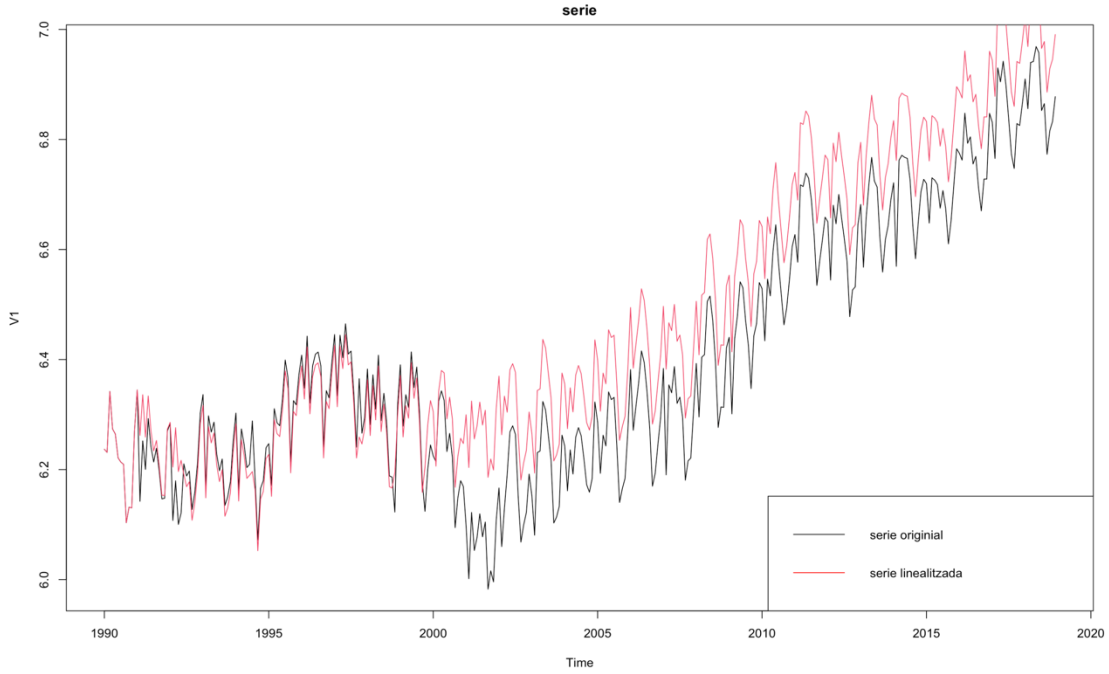
	Obs	type_detected	W_coeff	ABS_L_Ratio	Date	Efect
	<int>	<chr>	<dbl>	<dbl>	<chr>	<dbl>
2	14	TC	-0.11973938	3.578021	Feb 1991	88.71516
7	26	LS	-0.09539183	3.084992	Feb 1992	90.90167
3	30	LS	0.11507319	3.508742	Jun 1992	112.19556
12	55	AO	0.07250540	2.970973	Jul 1994	107.51986
5	94	AO	0.08646489	3.225424	Oct 1997	109.03131
4	107	AO	-0.09195310	3.379883	Nov 1998	91.21479
6	118	LS	-0.10030341	3.199818	Oct 1999	90.45629
11	122	TC	0.08902983	3.027707	Feb 2000	109.31133
1	133	LS	-0.12271595	3.610394	Ene 2001	88.45149
10	148	LS	0.09045023	3.044356	Abr 2002	109.46670
8	206	AO	-0.07918273	3.074737	Feb 2007	92.38711
9	290	AO	-0.07966472	3.118028	Feb 2014	92.34259

D'entrada s'observa que tots els atípics el que fan és incrementar el creixement sempre, mai decreixen. És lògic ja que des del 1990 l'únic que s'ha intentat, a tot el món, és incrementar l'ús d'energies renovables.

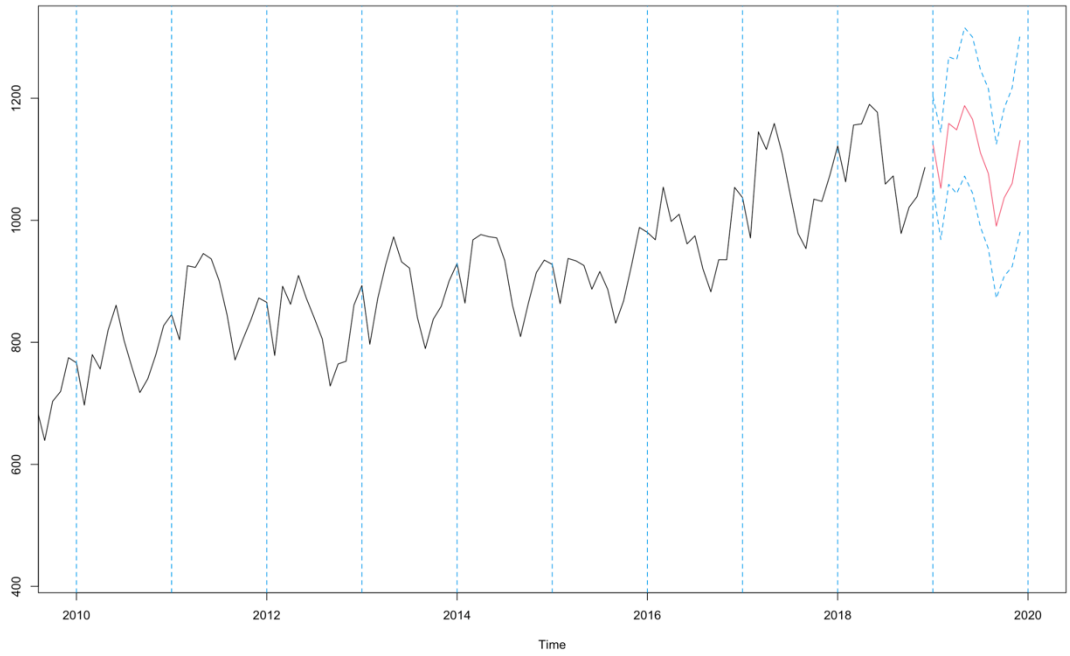
Tots aquells atípics que impliquen un canvi de nivell, és a dir, els de tipus LS (Level Shift), entenem que podrien ser degut a alguna llei aprovada pel govern nord-americà per a promoure l'ús d'energies renovables.

Els atípics que simplement impliquen un pic en un instant però immediatament retornen al seu nivell original, és a dir, els de tipus AO (athypic outlier), entenem que podria ser per un ús major d'energies renovables per raons com que ha plogut més i els embassaments estan més plens, que durant uns dies hi ha hagut dies molt intensos de sol i les plaques solars han produït molta energia...

Una vegada fet el tractament d'atípics dissenyem un nou model amb la sèrie linealitzada i la comparem amb la sèrie original per a observar si s'hi pot apreciar un canvi.



A més, realitzem una previsió per als següents 12 mesos, d'igual manera que hem fer en l'apartat anterior, però ara tenint en compte la sèrie linealitzada i dissenyant el model a partir d'aquesta.



A simple vista no s'observa gran canvi amb l'anterior però si es calcula la mitjana de l'amplada dels intervals de confiança ens dona un resultat menor (243,0607), 22 punts menor, és a dir, la variància del model ha disminuït després de fer el tractament d'atípics i linealitzar la sèrie.